The Astrophysical Journal, 609:1–14, 2004 July 1

APPLICATION OF MONTE CARLO ALGORITHMS TO THE BAYESIAN ANALYSIS OF THE COSMIC MICROWAVE BACKGROUND

J. Jewell, S. Levin, and C. H. Anderson
Jet Propulsion Laboratory, California Institute of Technology, MS 126-347, 4800 Oak Grove Drive, Pasadena, CA 91109-8099
Received 2002 August 29; accepted 2004 February 10

ABSTRACT

Power spectrum estimation and evaluation of associated errors in the presence of incomplete sky coverage; nonhomogeneous, correlated instrumental noise; and foreground emission are problems of central importance for the extraction of cosmological information from the cosmic microwave background (CMB). We develop a Monte Carlo approach for the maximum likelihood estimation of the power spectrum. The method is based on an identity for the Bayesian posterior as a marginalization over unknowns, and maximization of the posterior involves the computation of expectation values as a sample average from maps of the cosmic microwave background and foregrounds given some current estimate of the power spectrum or cosmological model, as well as some assumed statistical characterization of the foregrounds. Maps of the CMB and foregrounds are sampled by a linear transform of a Gaussian white-noise process, implemented numerically with conjugate gradient descent. For time series data with N_t samples and N pixels on the sphere, the method has a computational expense $KO(N^2 + N_t \log N_t)$, where K is a prefactor determined by the convergence rate of conjugate gradient descent. Preconditioners for conjugate gradient descent are given for scans close to great circle paths, and the method allows partial sky coverage for these cases by numerically marginalizing over the unobserved, or removed, region.

Subject headings: cosmic microwave background — methods: data analysis — methods: statistical

1. INTRODUCTION

Power spectrum estimation and evaluation of the associated errors in the presence of incomplete sky coverage; nonhomogeneous, correlated instrumental noise; and foreground emission are problems of central importance for the extraction of cosmological information from the cosmic microwave background (CMB). From a Bayesian point of view, power spectrum estimation involves the maximization of the posterior probability density, with error bars given by the set of cosmological parameters or power spectrum whose integrated posterior density achieves some specified level of confidence. A Bayesian approach to CMB analysis for large data sets involving a direct evaluation of the likelihood is intractable as a result of the $O(N^3)$ expense associated with computing the inverse of nonsparse matrices, or their determinants (Borrill 1999; Bond et al. 1999). The goal of this paper is the development of alternative numerical methods, specifically Monte Carlo techniques, for the Bayesian analysis of the CMB, including the complications of incomplete sky coverage, correlated noise, and foregrounds.

Previous work has demonstrated that for a certain class of scanning strategies, the signal and inverse noise matrices are block diagonal. The block diagonal properties of these matrices give an exact $O(N^2)$ Bayesian method, and therefore tractable for data sets as large as will be returned from *Planck*. The complications of this method are that it cannot easily accommodate partial sky converge or precessing scan strategies. The method of Oh et al. (1999) computes the maximum of the likelihood through a Newton-Raphson method. The numerical innovations of this method involve Monte Carlo simulations and the use of conjugate gradient descent, giving an overall expense $O(N^2)$. The method was proposed and numerically demonstrated in the context of uncorrelated noise and a region of sky coverage of azimuthal symmetry, where a good preconditioner can be constructed. However, the algorithm is in fact more general, provided that there is sufficient memory for storage of the needed matrices and that conjugate gradient descent converges quickly enough (i.e., there is a good preconditioner).

As suggested in Wandelt & Hansen (2003), we can use the ring set approach to supply preconditioners. An outstanding problem to be solved is a way of retaining the mathematical advantages of a ring set scan (block diagonal inverse noise and signal matrices) while accommodating partial sky coverage. The approach formulated in this paper handles the problem of partial sky coverage by embedding the data in an azimuthally symmetric region of sky and using a Monte Carlo Markov chain to numerically marginalize over the unobserved part. For scans close to ring sets, we therefore inherit good preconditioners, allowing an extension of both the ring set and conjugate gradient methods to scan strategies as planned for *Planck*.

For observations $d = s + f + \eta$, where (s, f, η) are the CMB signal, foregrounds, and noise, respectively, our approach to power spectrum estimation is motivated by the identity (derived in § A1)

$$\frac{p(\Gamma|d)}{p(\Gamma_0|d)} = \int d(s,f)\left[\frac{p(\Gamma|s)}{p(\Gamma_0|s)}\right]p(s,f|\Gamma_0,d), \qquad (1)$$

where Γ is any parameterization of conclusions (such as the power spectrum or cosmological parameters) and Γ_0 is any fixed guess. The Bayesian posterior ratio on the left is given as an integral over the unknown quantities that are assumed to generate the observed data. Maximization of the posterior involves computing the gradient of equation (1), which will be shown to depend on the expectation value of the power spectrum with respect to the random field $p(s,f|\Gamma_0,d)$,

$$E[\sigma_l(s)|\Gamma_0, d] = \int d(s,f)\sigma_l(s)p(s,f|\Gamma_0, d), \quad (2)$$

where we have defined the power at a given multipole order for a specific map

$$\sigma_l(s) = \frac{1}{2l+1}\sum_{-l\le m\le l} \|\langle lm|s\rangle\|^2 \quad (3)$$

(here and throughout the paper we use angle brackets to represent inner products). For this paper we choose the power spectrum itself as the parameterization and maximize the posterior ratio in equation (1) by the expectation maximization algorithm (Dempster et al. 1977), which proceeds by iteratively setting $C_l^{(n+1)} = E[\sigma_l(s)|C_l^{(n)}, d]$. The algorithm converges to the posterior maximum for a uniform prior and gives an unbiased, consistent estimator (see Appendix B). In this paper we focus on computation of the expectation value of the power spectrum $E[\sigma_l(s)|C_l^{(n)}, d]$ given the data and some guess $C_l^{(n)}$ under the assumption of perfect foreground separation (although we comment on how the approach can be generalized to include foregrounds later in the paper and leave its numerical demonstration for future work).

We compute the expectation value of the power spectrum numerically with a Monte Carlo approach, where we sample maps of the CMB from the probability density $p(s|f, d, C_l)$. Conditioning on some estimate of the foregrounds, the method exploits the fact that $p(s|f, C_l, d)$ is a Gaussian random field and therefore completely characterized by the mean field map and covariance matrix of fluctuations about that map. Maps are sampled from $p(s|f, C_l, d)$ by first computing the mean field map with conjugate gradient descent and then sampling fluctuations about the mean field map from a zero mean Gaussian field with covariance matrix $(N^{-1} + C^{-1})^{-1}$ (where N^{-1} is the inverse noise matrix and C^{-1} is the inverse covariance matrix for the CMB). These fluctuation maps are sampled by a linear transformation, numerically computed with conjugate gradient descent, of a spatial white-noise Gaussian process thereby generating maps with all the same statistical properties as samples from $p(s|f, C_l, d)$.

Each step of conjugate gradient descent involves a multiplication by the matrix $I + C^{1/2}N^{-1}C^{1/2}$, which can be done very quickly by multiplication by N^{-1} in the basis in which it is diagonal, followed by a transform to the spherical harmonic basis where C is diagonal. For spatially uncorrelated noise and circularly symmetric beams, we only need to transform from the pixel to the spherical harmonic domain, with an expense $O(N^{3/2})$ (Oh et al. 1999). In order to accommodate spatially correlated noise, we transform to the time domain, followed by a transform to the spherical harmonics, giving an expense $K[O(N^{3/2}) + N_t \log N_t]$, where N_t is the number of time samples and K depends on the convergence rate of conjugate gradient descent. Including the full complications of asymmetric beams, we would need to compute a convolution on the sphere. Using the convolution method of Wandelt & Gorski (2001), the expense of our method is $K[O(N^2) + N_t \log N_t]$.

The computational feasibility of this method is limited by finding a numerical implementation of conjugate gradient descent that converges quickly so that the prefactor K above is small. The strategy here is to embed the data in a region covered by an exact ring set scan, following the intuition that good preconditioners can be constructed for scan strategies close to ring sets (Wandelt & Hansen 2003). Embedding the data in a region on the sky with no observations (or where they have been removed) is accommodated by numerically marginalizing over the missing observations. Moreover, the same techniques can be used to marginalize over the foregrounds and provide Monte Carlo estimates of the confidence intervals for cosmological parameters.

The paper is organized as follows. We first review complications with a direct computation of the likelihood and provide an overview of our approach. We then discuss a technique we call "transformed white-noise sampling," which allows us to sample maps representing fluctuations about the mean field map for some guess of the power spectrum. We demonstrate the method with a flat sky 512×512 test case, including incomplete sky coverage, with uncorrelated, nonhomogeneous noise. We close with a discussion of further complications encountered in real CMB experiments and how they can be accommodated in the framework presented here.

2. POWER SPECTRUM ESTIMATION

2.1. *Likelihood*

We begin with a brief review of the likelihood and complications with its computational evaluation. The data returned from an experiment is a vector in the time domain $d(t)$, which is related to the CMB signal on the sky $s(n)$ through some linear mapping and additive Gaussian noise,

$$d(t) = \left\{\int dn'\, \delta[n(t) - n'] \int dn''\, B(n', n'')s(n'')\right\} + \eta(t), \quad (4)$$

where $B(n', n'')$ is the beam of the instrument and $\eta(t)$ is Gaussian noise, assumed to be stationary with a noise correlation matrix $N(t, t') = N(t - t')$. Denoting the linear mapping from sky to time domain as

$$R \equiv \int dn'\, \delta[n(t) - n'] \int dn''\, B(n', n''), \quad (5)$$

we simply write $d = Rs + \eta$. The likelihood for the power spectrum C_l given the data is

$$p(d|C_l) \propto \int ds\, e^{-(1/2)(d-Rs)N_t^{-1}(d-Rs)} e^{-(1/2)sC^{-1}s}. \quad (6)$$

Any linear transformation of the data vector will generate a Gaussian form for the likelihood. Transforming the data to an estimate of the CMB map $\hat{s} = (R^T N^{-1} R)^{-1} R^T N^{-1} d$ (as discussed in detail in Tegmark 1997; Stompor et al. 2002) with covariance matrix

$$E(\hat{s} \otimes \hat{s}) = C + \left(R^T N^{-1} R\right)^{-1} \quad (7)$$

gives the likelihood (up to an additive constant)

$$\log p[\hat{s}(d)|C_l] = -\tfrac{1}{2}\hat{s}(d)\left[\left(R^T N^{-1} R\right)^{-1} + C\right]^{-1}\hat{s}(d) - \tfrac{1}{2}\log\det\left[\left(R^T N^{-1} R\right)^{-1} + C\right]. \quad (8)$$

We can also write the likelihood up to an additive constant in the original time domain,

$$\log p(d|C_l) = -\tfrac{1}{2} d\left(N + RCR^T\right)^{-1} d - \tfrac{1}{2}\log\det\left(N + RCR^T\right). \quad (9)$$

Directly evaluating the likelihood in either the spatial or time domain results in an $O(N^3)$ computation, as it involves inversion, or computation of the determinant, of $[(R^T N^{-1} R)^{-1} + C]$ or $(N + RCR^T)$, respectively. The computational expense is due to the fact that we do not know the eigenbasis for either of these matrices, and computing this basis is generally an $O(N^3)$ problem.

The method of Oh et al. (1999) solves the likelihood in the spatial domain by evaluating the determinant with a Monte Carlo algorithm. The method involves conjugate gradient descent to solve a linear problem and as such involves matrix multiplication, which carries an $O(N^2)$ expense. The method was proposed and numerically demonstrated in the context of uncorrelated noise and a region of sky coverage of azimuthal symmetry. However, the algorithm is in fact more general, provided that there is sufficient memory for storage of the needed matrices and that conjugate gradient descent converges quickly enough (i.e., there is a good preconditioner).

For a certain class of observing strategies, we can exactly compute the likelihood, and for perturbations about these cases, we can use the approximate case as a preconditioner (as suggested in Wandelt & Hansen 2003). The approach formulated in this paper provides a consistent way to do this and involves a Monte Carlo Markov chain approach to numerically marginalizing over the unobserved part of azimuthally symmetric regions of the sky.

2.2. *Embedding the Data—Marginalization*

The method to be developed in this paper involves embedding the data in a region for which the signal and noise matrices have desirable properties. The likelihood for the data in the context of some model is given as an integral over the part of the embedding region that was not observed. This gives the identity for the Bayesian posterior for the power spectrum or cosmological model (denoted by Γ), given the time domain data $d(t)$ as the integral

$$\frac{p(\Gamma|d)}{p(\Gamma_0|d)} = \int d\left(s^{(1)},\ s^{(2)},\ f\right)\left[\frac{p\left(\Gamma|s^{(1)}, s^{(2)}\right)}{p(\Gamma_0|s^{(1)}, s^{(2)})}\right] \times p\left(s^{(1)}, s^{(2)},\ f|\Gamma_0, d\right), \quad (10)$$

where we have explicitly written the CMB maps $s = (s^{(1)}, s^{(2)})$ in terms of the part of the sky where we have data $s^{(1)}$ and the complementary region $s^{(2)}$. For the case of full sky coverage and prior knowledge $q(\Gamma)$, the log posterior ratio is given as

$$\log\frac{p\left(\Gamma|s^{(1)}, s^{(2)}\right)}{p(\Gamma_0|s^{(1)}, s^{(2)})} = \log\frac{q(\Gamma)}{q(\Gamma_0)} - \sum_l \left(l + \frac{1}{2}\right)\left\{\sigma_l\left(s^{(1)}, s^{(2)}\right) \times\left[\frac{1}{C_l(\Gamma)} - \frac{1}{C_l(\Gamma_0)}\right] + \log\frac{C_l(\Gamma)}{C_l(\Gamma_0)}\right\}, \quad (11)$$

where the power at a given multipole order is

$$\sigma_l\left(s^{(1)}, s^{(2)}\right) = \frac{1}{2l+1}\sum_{-l\le m\le l}\left\|\left\langle lm|s^{(1)} + s^{(2)}\right\rangle\right\|^2. \quad (12)$$

For other regions with azimuthal symmetry (an annulus or polar cap), the posterior ratio would involve a similar form in terms of block diagonal matrices. The simple form of the log posterior ratio for full sky coverage is one of the motivations for treating the problem of partial sky coverage as missing data and marginalizing over it (detailed and justified in § A2).

2.3. *Maximization*

In order to estimate the power spectrum, we would like to find Γ, which maximizes the posterior given the noisy data. Differentiating the posterior with respect to the parameters C_l gives a gradient in the direction

$$\left|\frac{\partial \log p(\Gamma|d)}{\partial C_l}\right|_{\Gamma_0} \propto \sum_l \left(l + \frac{1}{2}\right)\left[\frac{E(\sigma_l|d, \Gamma_0)}{C_l^2(\Gamma_0)} - \frac{1}{C_l(\Gamma_0)}\right], \quad (13)$$

where $E[C_l(s)|\Gamma_0, d]$ was given previously in equation (2). An improvement of our current estimate of the power spectrum Γ_0 can then be given according to a Newton-Raphson iterative scheme (Bond et al. 1998; Oh et al. 1999), where our current guess is updated using an approximation to the curvature of the likelihood $\boldsymbol{F}_{ll'}^{-1}(d, \Gamma_n)$, according to

$$C_l(\Gamma_{n+1}) = C_l(\Gamma_n) - \frac{1}{2}\sum_{l'}\boldsymbol{F}_{ll'}^{-1}(d, \Gamma_n)\left|\frac{\partial \log p(\Gamma|d)}{\partial C_l}\right|_{\Gamma_n}. \quad (14)$$

In Appendix B, it is shown that the curvature matrix is given in terms of the expectation value

$$E(\sigma_l\sigma_{l'}|d, \Gamma_0) = \int ds\, \sigma_l(s)\sigma_{l'}(s)p(s|d, \Gamma_0). \quad (15)$$

In practice, we might want to avoid computing the inverse of the curvature matrix and simply use the diagonal elements.

For this paper, we instead implemented the simpler (although more slowly converging) expectation maximization algorithm (Dempster et al. 1977). This method essentially follows from Jensen's inequality, giving the lower bound to the posterior

$$\log\left[\frac{p(\Gamma|d)}{p(\Gamma_0|d)}\right] \ge \log\frac{q(\Gamma)}{q(\Gamma_0)} - \sum_l\left(l + \frac{1}{2}\right)\left\{E[\sigma_l(s)|\Gamma_0, d] \times\left[\frac{1}{C_l(\Gamma)} - \frac{1}{C_l(\Gamma_0)}\right] + \log\frac{C_l(\Gamma)}{C_l(\Gamma_0)}\right\}. \quad (16)$$

For a uniform prior, the lower bound is maximized by $E[C_l(s)|\Gamma_0, d]$. In Appendix B, we prove that this estimator iteratively converges to the maximum of the posterior and is a consistent and unbiased estimator (for a uniform prior).

2.4. *Computing Expectation Values for the Power Spectrum*

In order to iteratively converge to the optimal, consistent estimator of the power spectrum, we need to compute, for any current guess of the power spectrum Γ_0, the expectation value $E[C_l(s)|d, \Gamma_0]$ as defined in equation (2). Defining the mean field map

$$\hat{s} = \left[N^{-1} + C^{-1}(\Gamma_0)\right]^{-1}N^{-1}d \quad (17)$$

and associated power spectrum estimate

$$\sigma_l(\hat{s}) = \frac{1}{2l+1}\sum_m\|\langle\hat{s}|lm\rangle\|^2, \quad (18)$$

the expectation value can be written by integrating over fluctuations about the mean field map $\xi = s - \hat{s}$ as

$$E[\sigma_l(s)|d,\Gamma_n] = \frac{1}{2l+1}\sum_m \int d\xi\, \langle \hat{s}+\xi|lm\rangle\langle lm|\hat{s}+\xi\rangle \times \frac{e^{-(1/2)\xi\left[N^{-1}+C^{-1}(\Gamma_n)\right]\xi}}{\int d\xi'\, e^{-(1/2)\xi'[N^{-1}+C^{-1}(\Gamma_n)]\xi'}}. \quad (19)$$

Since $E(\langle lm|\xi\rangle) = 0$, the expectation value is

$$E[\sigma_l(s)|d,\Gamma_n] = \sigma_l(\hat{s}) + \frac{1}{2l+1}\sum_m \int d\xi\, \langle \xi|lm\rangle\langle lm|\xi\rangle \times \frac{e^{-(1/2)\xi\left[N^{-1}+C^{-1}(\Gamma_n)\right]\xi}}{\int d\xi'\, e^{-(1/2)\xi'[N^{-1}+C^{-1}(\Gamma_n)]\xi'}}. \quad (20)$$

We refer to these two terms as the mean field map power spectrum estimate (known to be biased) and the correction term. Analytically, we know that the correction term is given by $(2l+1)^{-1}\sum_m \langle lm|[N^{-1}+C^{-1}(\Gamma_n)]^{-1}|lm\rangle$; however, it is intractable to compute and store the matrix $[N^{-1}+C^{-1}(\Gamma_n)]^{-1}$. Our strategy is to compute the correction term with a Monte Carlo method described below.

2.5. *Transformed White-Noise Sampling*

Because of the computational intractability of computing the matrix inverse $(N^{-1}+C^{-1})^{-1}$, our strategy is to compute the expectation value of the correction term from fluctuation maps ξ sampled from the zero mean Gaussian random field with covariance matrix $(N^{-1}+C^{-1})^{-1}$. We could easily sample fluctuation maps ξ if we could compute the eigenvectors and eigenvalues of the matrix $(N^{-1}+C^{-1})$, since in this basis the Gaussian probability density for ξ factors. However, computing the eigenvectors and eigenvalues is again an $O(N^3)$ operation.

Because of these difficulties, we look for an alternative way to sample maps. Defining $\delta = (N^{-1}+C^{-1})\xi$, we can write the log density, up to the normalization constant, as

$$-\xi\left(N^{-1}+C^{-1}\right)\xi = -\delta\left(N^{-1}+C^{-1}\right)^{-1}\delta. \quad (21)$$

The transformed Gaussian process has the covariance matrix $N^{-1}+C^{-1}$, making it easy to sample from. Specifically, we can sample maps from this Gaussian process by drawing two independent white-noise maps (ω_1, ω_2) and setting $\delta = C^{-1/2}\omega_1 + N^{-1/2}\omega_2$. Since both white-noise maps are drawn independently from a zero mean Gaussian process, the resulting covariance matrix is $E(\delta\otimes\delta) = N^{-1}+C^{-1}$ (as discussed in Appendix C).

The maps with the correct statistical properties are $\xi = (N^{-1}+C^{-1})^{-1}\delta$, which can be solved numerically for a given map δ. A numerically stable implementation (as also noted in Oh et al. 1999) involves conjugate gradient descent to solve for the maps $C^{-1/2}\xi$ (which can easily be transformed back to the fluctuations maps ξ),

$$\left(I + C^{1/2}N^{-1}C^{1/2}\right)\left(C^{-1/2}\xi\right) = \omega_1 + C^{1/2}N^{-1/2}\omega_2. \quad (22)$$

The resulting maps ξ have the correct statistical properties, since $E(\xi\otimes\xi) = (N^{-1}+C^{-1})^{-1}$ (see Appendix C), allowing us to compute the correction term to the power spectrum estimate of the mean field map as a sample average.

In order to actually sample fluctuation maps by transforming a Gaussian white-noise process, we need to obtain the Cholesky decomposition of both the signal and noise matrices. If we have observations with uncorrelated noise on the sky, then $N^{-1/2}$ is known in the spatial domain. However, the scan strategy of the instrument will result in complicated correlations, so that computing $N^{-1/2}$ is intractable. The noise is simple in the time domain, which suggests that instead of choosing white-noise maps in the spatial domain, we draw from white-noise Gaussian processes in the time domain, where we know $N^{-1/2}$ in the Fourier basis, followed by a transformation to the sky, where we can operate with $C^{1/2}$.

For a realization of a white-noise process in the time domain τ and a white-noise map in the spatial domain ω, we can compute a fluctuation map according to

$$\left(I + C^{1/2}R^T N^{-1}RC^{1/2}\right)\left(C^{-1/2}\xi\right) = \omega + C^{1/2}R^T N^{-1/2}\tau, \quad (23)$$

where $N^{-1/2}$ is known in the Fourier basis associated with the time domain. In Appendix C, this procedure is justified with a proof that the covariance matrix of the fluctuation maps is $E(\xi\otimes\xi) = [C^{-1}+(R^T N_t^{-1}R)^{-1}]^{-1}$.

2.6. *Computational Expense*

The overall computational expense is fixed, for each iteration of the power spectrum estimate, by the expense of matrix multiplication and number of iterations needed to converge with conjugate gradient descent. In order to multiply by the matrix $C^{1/2}R^T N^{-1}RC^{1/2}$, we need to perform the following:

1. Multiply by $C^{1/2}$: this is a diagonal matrix in the spherical harmonic domain if the embedding region is the full sky.
2. Transform to the time domain with the matrix R.
3. Compute a time domain fast Fourier transform (FFT).
4. Multiply by N^{-1}; this is a diagonal matrix in the time domain Fourier basis.
5. Compute a time domain inverse FFT.
6. Transform back to the spatial domain with R^T.
7. Compute a spherical harmonic transform.
8. Multiply again by $C^{1/2}$ in the spherical harmonic domain.

For the case of circularly symmetric beams, the convolution with the beam is not needed when operating with the matrix R or its transpose, giving an expense $KO(N^{3/2}+N_t\log N_t)$, where N_t is the number of time samples and K is the prefactor related to the convergence rate of conjugate gradient descent. For cases in which the beam is not circularly symmetric, the convolution with the beam would have to be computed, increasing the expense to $KO(N^2+N_t\log N_t)$.

3. NUMERICAL EXAMPLE

The simulations presented here involve the assumption of spatially uncorrelated, but nonhomogeneous, noise, as shown by the top left panel of Figure 1. We also restrict the problem to power spectrum estimation from a small patch of sky and neglect curvature (and therefore work with discrete Fourier basis instead of spherical harmonics). Our goal with these numerical simulations is to demonstrate the approach in action. Future work will involve numerical implementations on the sphere.

A CMB power spectrum was generated using CMBfast (Seljak & Zaldarriaga 1996), followed by the creation of a full

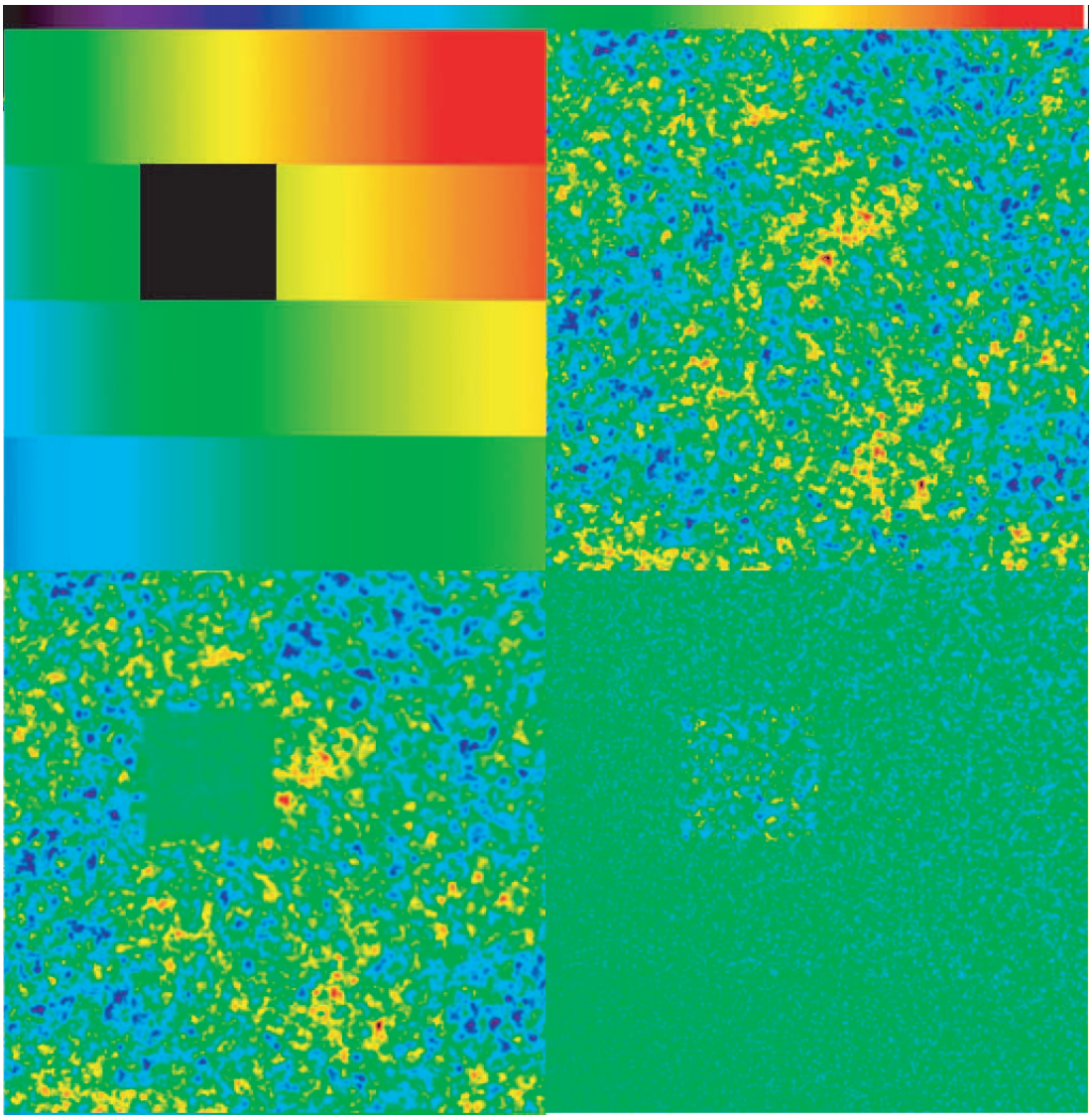

FIG. 1.—*Top left*: The rms of the noise in each pixel in units of the per pixel rms of the noise-free map, ranging from 0.5 in the lower left to 1.6 in the upper right. The black region is an unobserved "hole." *Top right*: Noise-free realization used as input: the map covers an area of sky $43^\circ \times 43^\circ$ with $5'$ pixels, smoothed with a $10'$ beam. *Bottom left*: Mean field map given a guess at the power spectrum and the simulated noisy data. *Bottom right*: Typical fluctuation map computed with conjugate gradient descent. A linear color bar is shown at the top of the figure.

sky map on the sphere using the SYNFAST routine in the HEALPIX package (Gorski et al. 1998[1]). A smaller patch of the sky was then selected and projected on a rectangular grid. This map $s(n)$ was taken to be the noise-free map, as shown in the top right panel of Figure 1. We then generated a noise map $\eta(n)$ by selecting independently at each pixel a Gaussian random number with variance scaled as shown according to the top left panel of Figure 1. Noise was added to the noise-free map and data then removed in a rectangular hole (as shown in Fig. 1). This was taken to be a simulated data set $d = s + \eta$ with partial coverage of the rectangular patch of sky.

The inverse noise matrix was given in terms of the variance at the ith pixel σ_i^2 as

$$N_{ij}^{-1} = \begin{cases} \delta_{ij}\sigma_i^{-2} & \text{for observed region of sky,} \\ 0 & \text{elsewhere.} \end{cases} \quad (24)$$

As an initial estimate of the power spectrum, we computed the power spectrum of the noisy, incomplete data (as computed in the two-dimensional Fourier basis since we neglected curvature) and subtracted the power spectrum of a single simulated noise map (on the region of sky where we have data). We then iteratively adjusted the power spectrum with the expectation maximization as above until convergence.

We found that preconditioning was in fact necessary to achieve a reasonable convergence rate. Using the initial estimate of the power spectrum $C(\Gamma_0)$ and computing the diagonal elements of the inverse noise matrix in the Fourier domain $\langle k|N^{-1}|k\rangle$ with Monte Carlo noise maps gave the preconditioner

$$M_{kk'}^{-1} \equiv \delta_{kk'}\left[I + C_{kk}^{1/2}(\Gamma_0)N_{kk}^{-1}C_{kk}^{1/2}\right]^{-1}. \quad (25)$$

Conjugate gradient descent was then used to solve the linear equations for the mean field and fluctuation maps

[1] See http://www.eso.org/science/healpix.

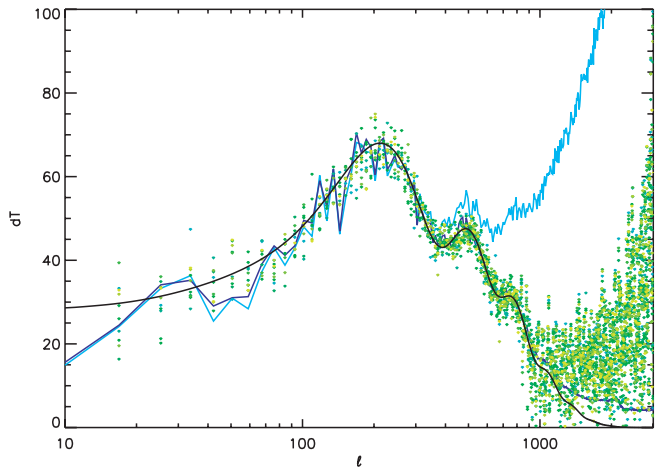

FIG. 2.—Plot showing the power spectrum estimation after 20 iterations starting from a flat initial guess. The green dots represent the results of separate runs on each of 10 simulated data sets, with each shade of green representing a different run. For each data set we have produced a new noise-free map (drawn from the theoretical power spectrum given by the solid black line) and added inhomogenous noise and a hole as shown in Fig. 1. The initial guess used was the expected (flat) noise spectrum, multiplied by a random number between 0 and 1. At low l the signal-to-noise ratio is high, and the spread in the dots is caused by noise and sample variance. At high l, the signal-to-noise ratio is low and the spread in the dots is caused by variation in the initial guess, resulting in an upper bound as shown in the plot. Also shown are sample power spectra for a single simulated data set (*light blue line*) and noise-free map (*dark blue line*).

$$M^{-1}\left(I + CN^{-1}\right)\hat{s} = M^{-1}CN^{-1}d,$$
$$M^{-1}\left(I + CN^{-1}\right)\xi = M^{-1}\delta, \tag{26}$$

where $\delta = C^{+1/2}\omega_1 + CN^{-1/2}\omega_2$ was computed from two independently chosen spatial white-noise maps (ω_1, ω_2) and $N^{-1/2}$ vanished in the unobserved part of the sky and elsewhere given by σ_i^{-1}. The result of iterating the algorithm to convergence is shown in Figure 2. Uncertainties in the power spectrum estimate were computed by Monte Carlo, in which new CMB maps were generated, noise added, and the algorithm run again.

4. ADDITIONAL PROBLEMS

The methods presented above can be generalized to handle other additional problems faced in the Bayesian analysis of the CMB. We do not provide numerical examples but briefly include comments on how to use transformed white-noise sampling to estimate error bars and include foregrounds.

4.1. *Confidence Intervals from a Markov Chain*

The ability to sample maps of the CMB given some estimate of the power spectrum can be used to construct a Markov chain Monte Carlo algorithm that converges to the Bayesian posterior $p(\Gamma|d)$ itself. Previously, Markov chain Monte Carlo techniques have been proposed for the extraction of marginal densities for cosmological parameters from approximate Bayesian posterior densities for the power spectrum (Christensen et al. 2001; Knox et al. 2001; Lewis & Bridle 2002; Runbino-Martin et al. 2003).

One way to derive a Markov chain algorithm that converges to the Bayesian posterior is to start directly from

$$p(\Gamma|d) = \int ds\, p(\Gamma, s|d). \tag{27}$$

Then using the fact that $p(\Gamma|s,d) = p(\Gamma|s)$ as discussed in § A1, we can write

$$\begin{aligned} p(\Gamma|d) &= \int ds\, p(\Gamma|s,d)p(s|d) \\ &= \int ds\, p(\Gamma|s)\, p(s|d) \\ &= \int ds\, p(\Gamma|s) \int d\Gamma'\, p(s,\Gamma'|d) \\ &= \int ds\, p(\Gamma|s) \int d\Gamma'\, p(s|\Gamma',d)p(\Gamma'|d) \\ &= \int d\Gamma' \left[\int ds\, p(\Gamma|s)p(s|\Gamma',d)\right] p(\Gamma'|d). \end{aligned} \tag{28}$$

This shows that the Bayesian posterior is the stationary density of the transition matrix

$$T(\Gamma|\Gamma';d) = \int ds\, p(\Gamma|s)p(s|\Gamma',d). \tag{29}$$

Therefore, sampling from any initial approximation to the Bayesian posterior and repeatedly sampling maps given parameter estimates, followed by variations in the parameters given the maps, gives the sequence of approximations

$$p_N(\Gamma|d) = \int d\Gamma' \left[\int ds\, p(\Gamma|s)p(s|\Gamma',d)\right] p_{N-1}(\Gamma'|d), \tag{30}$$

which will give convergence to the Bayesian posterior itself $p(\Gamma|d) = \lim_{N\to\infty} p_N(\Gamma|d)$.

The above can also be understood within the Metropolis-Hastings algorithm, which is a general framework for the construction of Markov chains that converge to a target stationary density. The goal is to construct a transition matrix for the *joint* density of CMB maps and the power spectrum: this can be done by first assuming detailed balance

$$p(\Gamma_1, s_1|d)T(\Gamma_2, s_2|\Gamma_1, s_1; d) = T(\Gamma_1, s_1|\Gamma_2, s_2; d)p(\Gamma_2, s_2|d). \tag{31}$$

From the condition of detailed balance we see that

$$p(\Gamma|d) = \int d(s, \Gamma_2, s_2)T(\Gamma, s|\Gamma_2, s_2; d)p(\Gamma_2, s_2|d), \tag{32}$$

which shows that the Bayesian posterior is the marginalized equilibrium density, generated by repeatedly taking steps generated with the transition matrix $T(\Gamma_2, s_2|\Gamma_1, s_1; d)$. Given any approximation to the joint density $p(\Gamma_2, s_2|d)$, repeated application of the transition matrix will reach the equilibrium density.

There are many ways to construct a transition matrix that has the desired stationary density (the fixed point of iterating with the transition matrix). Specifically, we are free to choose *any* proposal density $\rho(\Gamma_2, s_2|\Gamma_1, s_1; d)$, which is then related to the transition matrix through an acceptance or rejection of the proposed move. For any $\Gamma_2, s_2 \neq \Gamma_1, s_1$ the transition matrix is related to the proposal density according to

$$T(\Gamma_2, s_2|\Gamma_1, s_1; d) = a(\Gamma_2, s_2|\Gamma_1, s_1; d)\rho(\Gamma_2, s_2|\Gamma_1, s_1; d) \tag{33}$$

for any $\Gamma_2, s_2 \neq \Gamma_1, s_1$. A common choice for the acceptance probability, consistent with the condition of detailed balance, is

$$a(\Gamma_2, s_2|\Gamma_1, s_1; d) = \min\left[1, \frac{p(\Gamma_2, s_2|d)}{p(\Gamma_1, s_1|d)} \frac{\rho(\Gamma_1, s_1|\Gamma_2, s_2; d)}{\rho(\Gamma_2, s_2|\Gamma_1, s_1; d)}\right]. \tag{34}$$

Note that the unknown (and uncomputable) normalization constants have dropped out, and the acceptance probability can be explicitly computed for various choices of the proposal density. Specifically, the joint density $p(\Gamma, s, d)$ gives the ratio

$$\frac{p(\Gamma_2, s_2|d)}{p(\Gamma_1, s_1|d)} = \frac{p(\Gamma_2)}{p(\Gamma_1)} \times \frac{e^{-(1/2)(d-Rs_2)N^{-1}(d-Rs_2)-(1/2)s_2C^{-1}(\Gamma_2)s_2-\log\|C(\Gamma_2)\|}}{e^{-(1/2)(d-Rs_1)N^{-1}(d-Rs_1)-(1/2)s_1C^{-1}(\Gamma_1)s_1-\log\|C(\Gamma_1)\|}}. \tag{35}$$

One example of a proposal density can be given by alternately proposing a new map while holding the parameters fixed, and then in a second step varying the parameters while holding the maps fixed. Moreover, proposing these variations according to the exact conditional densities inherited from the full joint density $p(\Gamma, s, d)$ gives an acceptance probability of unity. To show this, we can set the proposal density for the first step of varying the map alone to

$$\rho(\Gamma_2, s_2|\Gamma_1, s_1; d) = \begin{cases} p(s_2|\Gamma_1; d) & \text{for } \Gamma_2 \equiv \Gamma_1, \\ 0 & \text{for } \Gamma_2 \neq \Gamma_1, \end{cases} \tag{36}$$

which gives the acceptance probability

$$\begin{aligned} a(\Gamma_2, s_2|\Gamma_1, s_1; d) &= \min\left[1, \frac{p(\Gamma_2, s_2|d)}{p(\Gamma_1, s_1|d)} \frac{\rho(\Gamma_1, s_1|\Gamma_2, s_2; d)}{\rho(\Gamma_2, s_2|\Gamma_1, s_1; d)}\right] \\ &= \min\left[1, \frac{p(s_2|d, \Gamma_1)p(\Gamma_1|d)}{p(s_1|d, \Gamma_1)p(\Gamma_1|d)} \frac{p(s_1|d, \Gamma_1)}{p(s_2|d, \Gamma_1)}\right] \\ &= 1, \end{aligned} \tag{37}$$

where in the second line above we have used $p(\Gamma, s|d) = p(s|\Gamma, d)p(\Gamma|d)$ and also the fact that $\Gamma_2 = \Gamma_1$. Similarly, in the second stage, we vary the parameters while holding the map fixed, so that

$$\rho(\Gamma_2, s_2|\Gamma_1, s_1; d) = \begin{cases} p(\Gamma_2|s_1) & \text{for } s_2 \equiv s_1, \\ 0 & \text{for } s_2 \neq s_1 \end{cases} \tag{38}$$

gives an acceptance matrix

$$\begin{aligned} a(\Gamma_2, s_2|\Gamma_1, s_1; d) &= \max\left[1, \frac{p(\Gamma_2, s_2|d)}{p(\Gamma_1, s_1|d)} \frac{\rho(\Gamma_1, s_1|\Gamma_2, s_2; d)}{\rho(\Gamma_2, s_2|\Gamma_1, s_1; d)}\right] \\ &= \max\left[1, \frac{p(\Gamma_2|s_1)p(s_1|d)}{p(\Gamma_1|s_1)p(s_1|d)} \frac{p(\Gamma_1|s_1)}{p(\Gamma_2|s_1)}\right] \\ &= 1, \end{aligned} \tag{39}$$

where in the second line we have used $p(\Gamma, s|d) = p(\Gamma|s)p(s|d)$ [which follows since $p(\Gamma|s, d) = p(\Gamma|s)$] and substituted $s_2 = s_1$. This example of a two-step proposal process is simply another way to understand, within the context of the Metropolis-Hastings algorithm, the identity for the Bayesian posterior given in equation (30).

In summary, a single step of the Markov chain, involving a transition from $(\Gamma_1, s_1) \rightarrow (\Gamma_2, s_2)$, involves the following:

1. Choose a new guess for the parameters from $p(\Gamma_2|s_1)$.
2. Sample a map from $p(s_2|d, \Gamma_2)$ according to the following:
   A. Compute the mean field map $\hat{s}_2 = [R^T N^{-1} R + C^{-1}(\Gamma_2)]^{-1} R^T N^{-1} d$.
   B. For two independently sampled white-noise maps in the spatial and time domains (ω, τ), compute ξ as the solution to $(I + C^{1/2} R^T N^{-1} R C^{1/2})(C^{-1/2}\xi) = \omega + C^{1/2} R^T N^{-1/2} \tau$.
   C. Set $s_2 = \hat{s}_2 + \xi$.
3. Continue.

For circularly symmetric beams, each step of the Markov chain has expense $KO(N^{3/2} + N_t \log N_t)$. While beyond the scope of this paper, future work will study the convergence rate of the Markov chain to the stationary density for other proposal densities.

4.2. *Inclusion of Foregrounds*

We conclude with a brief discussion of the inclusion of foregrounds. The inclusion of the foregrounds involves sampling both CMB and foreground maps given some estimate of the CMB power spectrum and some prior for the foregrounds. Examples of proposed foreground priors for a Bayesian treatment of their separation from the CMB include the maximum entropy method (Hobson et al. 1998) and Wiener filtering (Tegmark & Efstathiou 1996). Wiener filtering follows directly from the assumption that the prior for the foregrounds is Gaussian, with a known power spectrum. It should be noted that, although the foregrounds themselves are non-Gaussian, the separation of the foregrounds from CMB is driven not by the prior but by the quality of the data. We do not presume to have a complete (non-Gaussian) characterization of the foregrounds, but merely *some* information as to their statistical properties. The more that is known about the statistical structure of a particular source of emission, the easier it will be for it to be distinguished from other sources. Even knowing only the power spectrum of the foregrounds can improve the separation of the CMB and foregrounds (when included with data of sufficient *spectral* coverage), and it is this example we discuss below. While we restrict discussion to the Gaussian foreground priors, it should be noted that other foreground priors can be included within the numerical framework presented here, and this is one of the advantages of the approach.

Including foregrounds, the data for the jth frequency channel are given by

$$d_j(t) = A_{j0} R_j s + \sum_p A_{jp} R_j f_p + \eta_j, \tag{40}$$

where R_j is the mapping from the sky to the time domain for the jth frequency channel, A_{j0} is the response of the CMB at the jth frequency, A_{jp} is the response of the pth foreground component at the jth frequency, and we sum over the foreground components f_p. For a statistical characterization of the foregrounds β and a guess of the CMB power spectrum Γ, we need to sample from the density given by (up to normalization)

$$\begin{aligned} -\log p(s, f|d, \Gamma_0, \beta) \sim (d - A_0 Rs - ARf) N^{-1} (d - A_0 Rs \\ - ARf) - sC^{-1}(\Gamma_0)s - \log p(f|\beta). \end{aligned} \tag{41}$$

We need to compute the expectation value

$$E(C_l|d,\Gamma_0,\beta) = \int d(s,f)C_l(s)p(s,f|d,\Gamma_0,\beta). \quad (42)$$

This expectation value can be numerically computed by sampling maps (s,f) by the time average of a Markov chain with equilibrium density $p(s,f|d,\Gamma_0,\beta)$. One legitimate way to construct such a Markov chain is to alternately sample maps from the conditional densities $p(s|f,d,\Gamma_0,\beta)$ and $p(f|s,d,\Gamma_0,\beta)$. We briefly comment on sampling maps from each of these conditional densities.

Given some estimate of the foregrounds, we need to sample maps from the conditional density

$$-\log p(s|f,d,\Gamma_0,\beta) \sim (d - A_0Rs - ARf)N^{-1} \times (d - A_0Rs - ARf) - sC^{-1}(\Gamma_0)s. \quad (43)$$

Sampling from $p(s|f,d,\Gamma_0,\beta)$ proceeds in essentially the same way as described above, but generalized for multifrequency data. Specifically, the mean field map includes the subtracted response from the foreground estimate

$$\left(C^{-1} + R^TA_0^TN^{-1}A_0R\right)\hat{s} = R^TA_0^TN^{-1}(d - ARf) \quad (44)$$

and the fluctuation maps include a time domain white-noise sample for each frequency channel

$$\left(C^{-1} + R^TA_0^TN^{-1}A_0R\right)\xi = C^{-1/2}\omega_1 + \sum_j R_j^TA_{0j}N_j^{-1/2}\tau_j. \quad (45)$$

Note that the covariance matrix of the fluctuations is

$$E[\xi\otimes\xi] = \left(C^{-1} + R^TA_0^TN^{-1}A_0R\right)^{-1}, \quad (46)$$

where the proof depends on the independence of the time domain white-noise maps for every frequency channel.

Given some estimate of the CMB, we need to sample from the conditional density

$$-\log p(f|s,d,\Gamma_0,\beta) \sim (d - A_0Rs - ARf) \times N^{-1}(d - A_0Rs - ARf) - \log p(f|\beta). \quad (47)$$

If the prior for the foregrounds $p(f|\beta)$ is Gaussian,

$$-\log p(f|\beta) \sim fB^{-1}(\beta)f, \quad (48)$$

then we can also use transformed white-noise sampling. First, we compute the mean field foreground map,

$$\left(B^{-1} + R^TA_p^TN^{-1}A_pR\right)\hat{f} = R^TA_p^TN^{-1}(d - A_0Rs), \quad (49)$$

and then sampling fluctuations ξ_p according to the transform of white-noise samples

$$\left(B^{-1} + R^TA_p^TN^{-1}A_pR\right)\xi_p = B^{-1/2}\omega_1 + \sum_j R_j^TA_{pj}N_j^{-1/2}\tau_j. \quad (50)$$

Multiplying by the foreground signal matrix B is done in either the pixel or spherical harmonic basis depending on the basis in which it is sparse.

If we use a non-Gaussian prior (such as the maximum entropy method or other priors), then we will need to employ more general sampling techniques to sample foreground maps, such as the Metropolis algorithm. In this case the needed expectation value is to be computed as the *time average* from a Markov chain with equilibrium density $p(s,f|d,\Gamma_n)$.

5. CONCLUSIONS

The fundamental hurdle to numerically implementing an exact Bayesian approach to CMB analysis, including complications of partial sky coverage, correlated noise, and foregrounds, is finding efficient ways to solve the linear problem $(I + C^{1/2}R^TN^{-1}RC^{1/2})(C^{-1/2}\xi) = C^{1/2}\delta$ for any vector δ. Solving the linear equation has an expense $KO(N^{3/2} + N_t\log N_t)$ for circularly symmetric beams, and the algorithm provides a tractable approach provided that K can be made small enough. The strategy we presented in this paper allows the data to be embedded in an azimuthally symmetric region of the sky covered by a Wandelt ring set scan, with the intuition that, provided that the true scan of the instrument is close enough to the exact scan, we inherit good preconditioners (see Appendix D).

We also commented on how the method of transformed white-noise sampling can be used in Monte Carlo Markov chain for the entire Bayesian posterior. The feasibility of this approach depends on a good approximation to the posterior itself. Previous work has demonstrated several computationally feasible, unbiased estimates of the power spectrum and associated error covariance matrix. Any of these methods could therefore be used, in principle, to give an approximate posterior, so that a Markov chain approach can be used as a final consistency check.

Future work will incorporate the foregrounds in the algorithm presented here, generalized for multifrequency data. Maximization of the likelihood of the power spectrum given the data again leads to the computation of the expectation value $E(C_l|d,\Gamma_0)$, but now the marginalization includes the foregrounds as well. If the prior for the foregrounds is Gaussian, then we can also use transformed white-noise sampling to sample a new foreground map while conditioning on the CMB. If the prior used is non-Gaussian, other sampling schemes can be used, including Gibbs sampling or the Metropolis algorithm.

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contact with NASA. We also thank Eric Hivon and Ben Wandelt for interesting discussions during the course of this work.

APPENDIX A

IDENTITIES FOR THE BAYESIAN POSTERIOR

A1. IDENTITIES FOR THE POSTERIOR

The data returned from a CMB experiment is a vector $d(t)$ in the time domain, generated from scanning the CMB signal $s(n)$ and foregrounds $f(n)$ on the sky and adding independent Gaussian noise. The Bayesian posterior is given directly as an integral over

unknown quantities by

$$p(\Gamma|d) = \int d(s,f)\, p(\Gamma, s, f|d). \tag{A1}$$

For the case of Gaussian random fields, this integral can be done analytically, but evaluation of the resulting likelihood leads to computationally intractable matrix manipulations. For any estimate Γ_0

$$p(\Gamma|d) = \int d(s,f) \left[\frac{p(\Gamma, s, f|d)}{p(\Gamma_0, s, f|d)}\right] p(\Gamma_0, s, f|d). \tag{A2}$$

Since we have

$$\frac{p(\Gamma, s, f|d)}{p(\Gamma_0, s, f|d)} = \frac{p(\Gamma|s, f, d)}{p(\Gamma_0|s, f, d)}, \tag{A3}$$

this becomes

$$p(\Gamma|d) = p(\Gamma_0|d) \int d(s,f) \left[\frac{p(\Gamma|s, f, d)}{p(\Gamma_0|s, f, d)}\right] p(s, f|\Gamma_0, d), \tag{A4}$$

where we have used $p(\Gamma_0, s, f|d) = p(s, f|\Gamma_0, d) p(\Gamma_0|d)$. By the assumed independence of the noise, the joint density over which we are marginalizing is

$$p(\Gamma, s, f|d) \propto p(d|s) p(s|\Gamma) p(f|\beta) p(\Gamma), \tag{A5}$$

where $p(\Gamma)$ is a prior for the parameters, $p(f|\beta)$ is a prior for the foregrounds (with β a parameterization of the statistical properties of the foregrounds, such as the power spectrum), and $p(d|s)$ is completely determined by the noise properties of the instrument, the beam, and scan strategy. Given some estimate of the noise-free CMB signal, the density for a new guess of the power spectrum is independent of the data, $p(\Gamma|s, f, y) = p(\Gamma|s)$, as shown by

$$\begin{aligned} p(\Gamma|s, f, d) &= \frac{p(d|s) p(s|\Gamma) p(f) q(\Gamma)}{\int d\Gamma' p(d|s) p(s|\Gamma') p(f) q(\Gamma')} \\ &= \frac{p(s|\Gamma) q(\Gamma)}{\int d\Gamma' p(s|\Gamma') q(\Gamma')}. \end{aligned} \tag{A6}$$

Therefore, for *any estimate* Γ_0, our identity now reads

$$\frac{p(\Gamma|d)}{p(\Gamma_0|d)} = \int d(s,f) \left[\frac{p(\Gamma|s)}{p(\Gamma_0|s)}\right] p(s, f|\Gamma_0, d), \tag{A7}$$

or for the likelihood ratio

$$\frac{p(d|\Gamma)}{p(d|\Gamma_0)} = \int d(s,f) \left[\frac{p(s|\Gamma)}{p(s|\Gamma_0)}\right] p(s, f|\Gamma_0, d). \tag{A8}$$

A2. PARTIAL SKY COVERAGE

The likelihood for the CMB given the theory for partial sky coverage can be written as the marginalization over the unobserved part of the sky. Denoting the CMB $s = (s^{(1)}, s^{(2)})$ as the CMB in the observed and unobserved regions of sky, respectively, we have

$$p\left(s^{(1)}|\Gamma\right) = \int ds^{(2)}\, p\left(s^{(1)}, s^{(2)}|\Gamma\right). \tag{A9}$$

Therefore, the posterior for partial sky coverage can be written

$$\begin{aligned} p(\Gamma|d) &\propto q(\Gamma) \int d\left(s^{(1)}, f\right) p\left(d|s^{(1)}, f\right) p\left(s^{(1)}|\Gamma\right) p(f|\beta) \\ &= q(\Gamma) \int d\left(s^{(1)}, f\right) p\left(d|s^{(1)}, f\right) \left[\int ds^{(2)}\, p\left(s^{(1)}, s^{(2)}|\Gamma\right)\right] p(f|\beta) \\ &= q(\Gamma) \int d\left(s^{(1)}, s^{(2)}, f\right) p\left(d|s^{(1)}, f\right) p\left(s^{(1)}, s^{(2)}|\Gamma\right) p(f|\beta). \end{aligned} \tag{A10}$$

This gives the identity, explicitly written for arbitrary sky coverage,

$$\frac{p(\Gamma|d)}{p(\Gamma_0|d)} = \int d\left(s^{(1)}, s^{(2)}, f\right)\left[\frac{p\left(\Gamma|s^{(1)}, s^{(2)}\right)}{p(\Gamma_0|s^{(1)}, s^{(2)})}\right] p\left(s^{(1)}, s^{(2)}, f|\Gamma_0, d\right). \tag{A11}$$

Because $s \equiv (s^{(1)}, s^{(2)})$ is supported on the full sky, the log posterior ratio is

$$\log\frac{p(\Gamma|s)}{p(\Gamma_0|s)} = \log\frac{q(\Gamma)}{q(\Gamma_0)} - \sum_l \left(l+\frac{1}{2}\right)\left\{C_l(s)\left[\frac{1}{C_l(\Gamma)} - \frac{1}{C_l(\Gamma_0)}\right] + \log\frac{C_l(\Gamma)}{C_l(\Gamma_0)}\right\} \tag{A12}$$

and the conditional density from which we are to sample $(s^{(1)}, s^{(2)})$ is

$$-\log p\left(s^{(1)}, s^{(2)}|\Gamma_0, d\right) \sim \tfrac{1}{2}\left(d - s^{(1)}\right)N^{-1}\left(d - s^{(1)}\right) + \tfrac{1}{2}\left(s^{(1)}, s^{(2)}\right)C^{-1}(\Gamma)\left(s^{(1)}, s^{(2)}\right). \tag{A13}$$

This is equivalent to setting the inverse noise matrix to zero for the part of the sky where there are no data.

A3. NO DATA AND NO NOISE LIMITS

Two limiting cases of our identity involve "no data" and "no noise." In the no-data limit, the inverse noise matrix vanishes everywhere, so that $p(s|y, \Gamma_0) \to p(s|\Gamma_0)$ and the posterior is given by the prior, as shown by

$$\begin{aligned}\frac{p(\Gamma|d)}{p(\Gamma_0|d)} &= \int ds\left[\frac{p(\Gamma|s)}{p(\Gamma_0|s)}\right]p(s|\Gamma_0)\\ &= \int ds\left[\frac{p(s|\Gamma)q(\Gamma)}{p(s|\Gamma_0)q(\Gamma_0)}\right]p(s|\Gamma_0)\\ &= \frac{q(\Gamma)}{q(\Gamma_0)}\int ds\, p(s|\Gamma)\\ &= \frac{q(\Gamma)}{q(\Gamma_0)}.\end{aligned} \tag{A14}$$

In the noise-free limit, the conditional density $p(s|y, \Gamma_0) \to \delta(s - s_{\rm true})$, independent of our choice of Γ_0, so that we converge to the noise-free posterior ratio.

APPENDIX B

PROPERTIES OF THE POWER SPECTRUM ESTIMATOR

B1. POSTERIOR MAXIMUM

Recall that the algorithm used involves iterating

$$C_l(\Gamma_{n+1}) = E(\sigma_l|\Gamma_n, d) \tag{B1}$$

and therefore the fixed point satisfies $C_l(\Gamma) = E(\sigma_l|\Gamma, d)$. We can prove that this estimator maximizes the posterior, or equivalently the log posterior. A direct calculation shows that

$$\frac{\partial}{\partial C_l(\Gamma)}\frac{p(\Gamma|d)}{p(\Gamma_0|d)} = \int ds\left\{\sum_l\left(l+\frac{1}{2}\right)\sigma_l(s)\left[\frac{1}{C_l^2(\Gamma)} - \frac{1}{C_l(\Gamma)}\right]\right\}\frac{p(\Gamma|s)}{p(\Gamma_0|s)}p(s|d, \Gamma_0), \tag{B2}$$

so that at the maximum

$$0 = \int ds\left\{\sum_l\left(l+\frac{1}{2}\right)\sigma_l(s)\left[\frac{1}{C_l^2(\Gamma)} - \frac{1}{C_l(\Gamma)}\right]\right\}p(s|d, \Gamma_0). \tag{B3}$$

At the maximum we therefore have

$$C_l(\Gamma_0) = \int ds\,\sigma_l(s)p(s|d, \Gamma_0), \tag{B4}$$

which is identically the fixed point of the expectation maximization algorithm. Therefore, the iteration converges to the maximum of the posterior (for a uniform prior).

B2. ESTIMATING THE CURVATURE OF THE LIKELIHOOD

After we have computed the maximum likelihood estimator of the power spectrum (or maximum posterior estimate), we want to find a confidence interval. There are several ways this confidence interval might be approximated. One approach is to compute the inverse curvature matrix of the likelihood and take the diagonal entries as an estimate of the error bars. We comment below on how this can be done with the same method used to compute $E(\sigma_l|d,\Gamma)$.

Approximating the likelihood as a Gaussian functional of the $C_l(\Gamma)$ is equivalent to a second-order Taylor expansion of the log likelihood about the maximum. As before, we have the identity for the likelihood

$$\frac{p(d|\Gamma)}{p(d|\Gamma_0)} = \int ds\, \frac{p(s|\Gamma)}{p(s|\Gamma_0)} p(s|d,\Gamma_0). \tag{B5}$$

It is more convenient to parameterize the likelihood ratio in terms of $\theta_l = C_l^{-1}$, so that when embedding the data on the full sky, we have

$$\log \frac{p(s|\Gamma)}{p(s|\Gamma_0)} = -\sum_l \left(l+\frac{1}{2}\right)\left\{[\theta_l(\Gamma)-\theta_l(\Gamma_0)]\sigma_l(s) - \log\frac{\theta_l(\Gamma)}{\theta_l(\Gamma_0)}\right\}. \tag{B6}$$

Denoting the curvature matrix of the likelihood

$$\boldsymbol{F}_{ll'}(d,\Gamma_0) \equiv -\frac{\partial^2 \log p(d|\Gamma)}{\partial C_l \partial C_{l'}}, \tag{B7}$$

we have the relation

$$-\boldsymbol{F}_{ll'}(d,\Gamma_0) = \frac{1}{C_l^2(\Gamma_0)} \frac{\partial^2 \log p(d|\Gamma)}{\partial\theta_l \partial\theta_{l'}} \frac{1}{C_{l'}^2(\Gamma_0)}. \tag{B8}$$

The curvature of $\log p(d|\Gamma)$ evaluated at the maximum, where $C_l(\Gamma_0) = E(C_l|d,\Gamma_0)$, is therefore

$$\frac{\partial^2 \log p(d|\Gamma)}{\partial\theta_l \partial\theta_{l'}} = -\delta_{ll'}\left(l+\tfrac{1}{2}\right)C_l^2(\Gamma_0) + \left(l+\tfrac{1}{2}\right)\left(l'+\tfrac{1}{2}\right)[E(\sigma_l\sigma_{l'}|d,\Gamma_0) - C_l(\Gamma_0)C_{l'}(\Gamma_0)], \tag{B9}$$

where we have defined the expectation value

$$E(\sigma_l\sigma_{l'}|d,\Gamma_0) = \int d\xi\, \sigma_l(\hat{s}+\xi)\sigma_{l'}(\hat{s}+\xi) \frac{e^{-(1/2)\xi\left[N^{-1}+C^{-1}(\Gamma_0)\right]\xi}}{\int d\xi'\, e^{-(1/2)\xi'[N^{-1}+C^{-1}(\Gamma_0)]\xi'}}. \tag{B10}$$

This can, in principle, be computed with the conjugate gradient descent method of transforming samples from a white-noise process. Future work will study the accuracy and convergence properties of estimating the curvature matrix from transformed white-noise sampling.

B3. COVARIANCE MATRICES

The correctness of the algorithm for power spectrum estimation presented in this paper is established by proving that the covariance matrix of two linearly transformed vectors has some specific form. For simplicity of notation, we choose to denote the covariance matrix of two vectors (x, y) as the expectation value of the outer product of the vectors $E(x_i y_j) \equiv E(x \otimes y)_{ij}$. For two matrices A and B, an identity used repeatedly in computing covariance matrices is

$$E[(Ax)\otimes(By)] = AE(x\otimes y)B^T, \tag{B11}$$

which is shown simply by checking for each matrix element

$$\begin{aligned} E[(Ax)\otimes(By)]_{mn} &= \sum_{ij} E\left(A_{mi}x_i B_{nj} y_j\right) \\ &= \sum_{ij} A_{mi} E\left(x_i y_j\right) B^T_{jn}. \end{aligned} \tag{B12}$$

One example of this identity is in computing the expectation value of maps computed from time-ordered data. One form of making a map from time-ordered data is given by

$$\hat{s} = \left(R^T N^{-1} R\right)^{-1} R^T N^{-1} d \tag{B13}$$

as mentioned in § 2 and in Tegmark (1997). According to the identity above, the covariance matrix of this map is

$$E(\hat{s} \otimes \hat{s}) = \left(R^T N^{-1} R\right)^{-1} R^T N^{-1} E(d \otimes d) \left(R^T N^{-1}\right)^T \left[\left(R^T N^{-1} R\right)^{-1}\right]^T. \tag{B14}$$

Substituting $d = Rs + \eta$, this is

$$\begin{aligned} E(\hat{s} \otimes \hat{s}) &= \left(R^T N^{-1} R\right)^{-1} R^T N^{-1} E[(Rs + \eta) \otimes (Rs + \eta)] N^{-1} R \left(R^T N^{-1} R\right)^{-1} \\ &= \left(R^T N^{-1} R\right)^{-1} R^T N^{-1} \left[RE(s \otimes s) R^T + RE(s \otimes \eta) + E(\eta \otimes s) R^T + E(\eta \otimes \eta)\right] N^{-1} R \left(R^T N^{-1} R\right)^{-1} \\ &= \left(R^T N^{-1} R\right)^{-1} R^T N^{-1} \left(RCR^T + N\right) N^{-1} R \left(R^T N^{-1} R\right)^{-1} \\ &= C + \left(R^T N^{-1} R\right)^{-1}, \end{aligned} \tag{B15}$$

where we have used the independence of the signal and noise and also that both are zero mean processes.

B4. CONSISTENCY

The estimator given by the expectation maximization algorithm (also equivalent to the maximum likelihood estimator) is given by

$$E(s \otimes s|d) = \int ds\, (s \otimes s) \frac{e^{-(1/2)(d-Rs)N^{-1}(d-Rs) - (1/2)sC^{-1}s}}{\int ds'\, e^{-(1/2)(d-Rs')N^{-1}(d-Rs') - (1/2)s'C^{-1}s'}}. \tag{B16}$$

For partial sky coverage, this is equivalent to embedding the data on the full sky and marginalizing over the "missing" observations

$$E(s \otimes s|d) = \int d\left(s, d^{(2)}\right) (s \otimes s) \frac{e^{-(1/2)(d^{(2)}-s)\tilde{N}^{-1}(d^{(2)}-s)} e^{-(1/2)(d-Rs)N^{-1}(d-Rs) - (1/2)sC^{-1}s}}{\int d(s', y')\, e^{-(1/2)(y'-s')\tilde{N}^{-1}(y'-s')} e^{-(1/2)(y-Rs')N^{-1}(y-Rs') - (1/2)s'C^{-1}s'}}, \tag{B17}$$

where we have arbitrarily chosen the full sky inverse noise matrix

$$N^{-1} = \begin{bmatrix} R^T N^{-1} R & 0 \\ 0 & \tilde{N}^{-1} \end{bmatrix}. \tag{B18}$$

The expectation of the covariance matrix is then

$$E[E(s \otimes s|d)] = \int d\mu(d) \int d\left(s, d^{(2)}\right) (s \otimes s) \frac{e^{-(1/2)(d^{(2)}-s)\tilde{N}^{-1}(d^{(2)}-s)} e^{-(1/2)(d-Rs)N^{-1}(d-Rs) - (1/2)sC^{-1}s}}{\int d(s', d')\, e^{-(1/2)(d'-s')\tilde{N}^{-1}(d'-s')} e^{-(1/2)(d-Rs')N^{-1}(d-Rs') - (1/2)s'C^{-1}s'}} \tag{B19}$$

[where we have indicated an integration over the data with the usual Lebesgue measure $d\mu(d)$ in order to avoid the confusing notation $d(d)$]. Consistency of the estimator is then shown by proving that $C = E[E(s \otimes s|d)]$.

Using the augmented noise matrix (which now has an inverse on the full sky), we can define the mean field map $\hat{s} = (N^{-1} + C^{-1})^{-1} N^{-1}(d, d^{(2)})$, with covariance matrix

$$\begin{aligned} E[E(\hat{s} \otimes \hat{s}|d)] &= \left(N^{-1} + C^{-1}\right)^{-1} N^{-1} E(d \otimes d) N^{-1} \left(N^{-1} + C^{-1}\right)^{-1} \\ &= C(N + C)^{-1} (N + C)(N + C)^{-1} C \\ &= C(N + C)^{-1} C. \end{aligned} \tag{B20}$$

The expectation value of the correction is $(N^{-1} + C^{-1})^{-1}$, so that the covariance matrix for the sum of the mean field and fluctuation maps is

$$\begin{aligned} E[E(s \otimes s|d)] &= C(N + C)^{-1} C + \left(N^{-1} + C^{-1}\right)^{-1} \\ &= C(N + C)^{-1} C + N(N + C)^{-1} C \\ &= C. \end{aligned} \tag{B21}$$

Therefore, the expectation maximization algorithm converges to the maximum likelihood estimator for a uniform prior, which is also a consistent estimator for arbitrary sky coverage.

APPENDIX C

CORRECTNESS OF TRANSFORMED WHITE-NOISE SAMPLING

As discussed in § 2.5, the algorithm to sample maps from the Gaussian process with covariance matrix $(N^{-1} + C^{-1})^{-1}$ is as follows:

1. Draw (ω_1, ω_2) from a white-noise process.
2. Compute $\delta = C^{-1/2}\omega_1 + N^{-1/2}\omega_2$.
3. Compute $\xi = (C^{-1} + N^{-1})^{-1}\delta$ with conjugate gradient descent.

We can prove that for any N^{-1}, even one that does not have an inverse (i.e., as in the case of partial coverage of the chosen embedding region), we have $E(\xi \otimes \xi) = (N^{-1} + C^{-1})^{-1}$. From the above we have

$$\xi = \left(N^{-1} + C^{-1}\right)^{-1}\left(C^{-1/2}\omega' + N^{-1/2}\omega\right), \tag{C1}$$

which gives

$$\begin{aligned} E(\xi \otimes \xi) &= \left(N^{-1} + C^{-1}\right)^{-1} E\left[\left(C^{-1/2}\omega' + N^{-1/2}\omega\right) \otimes \left(C^{-1/2}\omega' + N^{-1/2}\omega\right)\right]\left(N^{-1} + C^{-1}\right)^{-1} \\ &= \left(N^{-1} + C^{-1}\right)^{-1}\Big[C^{-1/2}E(\omega' \otimes \omega')C^{-1/2} + C^{-1/2}E(\omega' \otimes \omega)N^{-1/2} \\ &\quad + N^{-1/2}E(\omega \otimes \omega')C^{-1/2} + N^{-1/2}E(\omega \otimes \omega)N^{-1/2}\Big]\left(N^{-1} + C^{-1}\right)^{-1} \\ &= \left(N^{-1} + C^{-1}\right)^{-1}\left(C^{-1} + N^{-1}\right)\left(N^{-1} + C^{-1}\right)^{-1} \\ &= \left(N^{-1} + C^{-1}\right)^{-1}, \end{aligned} \tag{C2}$$

where by independence of the two white-noise maps $E(\omega \otimes \omega') = E(\omega) \otimes E(\omega')$, which vanishes since the white-noise process is zero mean. An important point to notice is that the matrices N^{-1} or $N^{-1/2}$ can be singular in the sense that they do not have inverses on the full sky (i.e., are generated by incomplete scanning of the sky). In fact, $N^{-1/2}\omega$ vanishes in the null space of N^{-1} (where we do not have data).

For a realization of a white-noise process in the time domain τ and a white-noise map in the spatial domain ω, we compute a fluctuation map according to

$$\xi = \left(C^{-1} + R^T N^{-1} R\right)^{-1}\left(C^{-1/2}\omega + R^T N^{-1/2}\tau\right), \tag{C3}$$

where $N^{-1/2}$ is known in the Fourier basis associated with the time domain. We can prove that the maps ξ have the correct covariance matrix, $E(\xi \otimes \xi) = (R^T N^{-1} R + C^{-1})^{-1}$, by the direct calculation

$$\begin{aligned} E(\delta \otimes \delta) &= E\left[\left(C^{-1/2}\omega + R^T N^{-1/2}\tau\right) \otimes \left(C^{-1/2}\omega + R^T N^{-1/2}\tau\right)\right] \\ &= \left[C^{-1/2}E(\omega \otimes \omega)C^{-1/2} + R^T N^{-1/2}E(\tau \otimes \tau)N^{-1/2}R\right] \\ &= \left(C^{-1} + R^T N^{-1} R\right), \end{aligned} \tag{C4}$$

where again the cross terms vanish since independence gives $E(\omega \otimes \tau) = E(\omega) \otimes E(\tau)$, which vanishes since both are zero mean processes. Then this gives the covariance matrix

$$\begin{aligned} E(\xi \otimes \xi) &= \left(R^T N^{-1} R + C^{-1}\right)^{-1} E(\delta \otimes \delta)\left(R^T N^{-1} R + C^{-1}\right)^{-1} \\ &= \left(R^T N^{-1} R + C^{-1}\right)^{-1}\left(C^{-1} + R^T N^{-1} R\right)\left(R^T N^{-1} + C^{-1}\right)^{-1} \\ &= \left(R^T N^{-1} R + C^{-1}\right)^{-1}, \end{aligned} \tag{C5}$$

so that the fluctuation maps do have the target covariance matrix.

APPENDIX D

EMBEDDING THE DATA IN AZIMUTHALLY SYMMETRIC REGIONS

We can also find the likelihood for the data embedded in an azimuthally symmetric region covered by a ring set. As discussed in Wandelt & Hansen (2003), we represent the signal on the ring set with coefficients $a_{mm'}$, so that

$$s(\theta,\phi)=\sum_{mm'}a_{mm'}e^{-im\theta}e^{-im'\phi}, \tag{D1}$$

where both indices range from $-L_{\max}\le m\le L_{\max}$. Therefore, the signal on the sky, parameterized on the ring set, is given by a two-dimensional inverse FFT. For any specific power spectrum, the corresponding signal matrix on the ring set covering the embedding region is block diagonal. We denote the ring set covariance matrix $T=E(a\otimes a)$.

Given the noise-free signal on the ring set $s(\theta,\phi)$, any observed data set is given by projecting into the subregion where we have data,

$$d(t)=Qs+\eta(t), \tag{D2}$$

where Q vanishes on the regions of the ring set where we have no observations. As before, we would first compute the mean field and fluctuation maps. In order to sample fluctuation maps, we would again compute

$$\left(I+TQ^TN^{-1}Q\right)\xi=T^{+1/2}\omega+\mathcal{C}Q^TN^{-1/2}\tau, \tag{D3}$$

where τ is a time domain white-noise process, ω is a spatial white-noise process on the full sky, and $T^{+1/2}$ is the Cholesky decomposition of the ring set covariance matrix.

Although it is possible to compute the Cholesky decomposition of the ring set covariance matrix, we might instead sample $C^{1/2}\omega$ on the full sky and then project into the region covered by the ring set, giving maps $RC^{1/2}\omega$. Then the covariance matrix for these maps is, using the usual identity,

$$\begin{aligned}E\left[\left(RC^{1/2}\omega\right)\otimes\left(RC^{1/2}\omega\right)\right]&=RC^{1/2}E(\omega\otimes\omega)C^{1/2}R^T\\&=RCR^T,\end{aligned} \tag{D4}$$

which is the correct covariance matrix on the ring set region. The point is that we do not have to compute the Cholesky decomposition of the signal matrix on the ring set but can instead transform white-noise maps with the Cholesky decomposition of the full sky signal covariance matrix (diagonal in the spherical harmonic basis), and then project down to the ring set region.

We can construct a preconditioner as follows. Define the projection from the time domain to the ring set following a Wandelt scan strategy as $W^TN^{-1}W$. For the same time domain noise matrix N^{-1}, we can use a preconditioner given by $W^TN^{-1}W$ and define

$$M^{-1}=\left(I+TW^TN^{-1}W\right)^{-1}. \tag{D5}$$

It is shown in Wandelt & Hansen (2003) that $W^TN^{-1}W$ is block diagonal on the ring set, so that M^{-1} can be computed in $O(N^2)$ operations. We can then solve the linear equations for the mean field and fluctuation maps

$$\begin{aligned}M^{-1}\left(I+TQ^TN^{-1}Q\right)\hat{s}&=M^{-1}TQ^TN^{-1}d,\\M^{-1}\left(I+TQ^TN^{-1}Q\right)\xi&=M^{-1}\left(T^{+1/2}\omega+TQ^TN^{-1/2}\tau\right).\end{aligned} \tag{D6}$$

For scans that are close to the ring set scan the intuition is that conjugate gradient descent will converge quickly for the above equations.

REFERENCES

Bond, J. R., Jaffe, A. H., & Knox, L. 1998, Phys. Rev. D, 57, 2117
Bond, J. R., et al. 1999, Comput. Sci. Eng., 1, 21
Borrill, J. 1999, Phys. Rev. D, 59, 027302
Christensen, N., Meyer, R., Knox, L., & Luey, B. 2001, Classical Quantum Gravity, 18, 2677
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. R. Stat. Soc. B, 39, 1
Gorski, E. K., Hivon, E., & Wandelt, B. 1998, in Proceedings of the MPA/ESO Garching Conference, ed. A. J. Banday, K. Sheth, & L. Da Costa (Garching: ESO), in press, http://www.eso.org/science/healpix
Hobson, M. P., et al. 1998, MNRAS, 300, 1
Knox, L., et al. 2001, ApJ, 563, L95
Lewis, A., & Bridle, S. 2002, Phys. Rev. D, 66, 103511
Oh, S. P., Spergel, D. N., & Hinshaw, G. 1999, ApJ, 510, 551
Runbino-Martin, R. A., et al. 2003, MNRAS, 341, 1084
Seljak, U., & Zaldarriaga, M. 1996, ApJ, 469, 437
Stompor, R., et al. 2002, Phys. Rev. D, 65, 022003
Tegmark, M. 1997, Phys. Rev. D, 55, 5895
Tegmark, M., & Efstathiou, G. 1996, MNRAS, 281, 1297
Wandelt, B., & Gorski, K. 2001, Phys. Rev, D, 63, 123002
Wandelt, B., & Hansen, F. 2003, Phys. Rev. D, 67, 023001